NYDP Shooting Incidents Over Time: An Analysis

```
# install.packages("tidyverse")
# install.packages("lubridate")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)

Shootings <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

```
## Rows: 27312 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Shootings$PRECINCT <- as.integer(Shootings$PRECINCT)

# Ensure each incident does not have a value of NA for several values:
Shootings <- Shootings[!(is.na(Shootings$OCCUR_DATE)),]
Shootings <- Shootings[!(is.na(Shootings$OCCUR_TIME)),]
Shootings <- Shootings[!(is.na(Shootings$PRECINCT)),]
Shootings <- Shootings[!(is.na(Shootings$BORO)),]
```

One of the largest problems facing the NYPD today is staffing shortages:

- https://www.cbsnews.com/newyork/news/independent-budget-office-data-reveals-nypd-is-down-1200-members-from-2022-and-2900-from-2019/
- https://nypost.com/2023/03/10/nypd-cops-resigning-from-force-in-2023-at-record-pace/

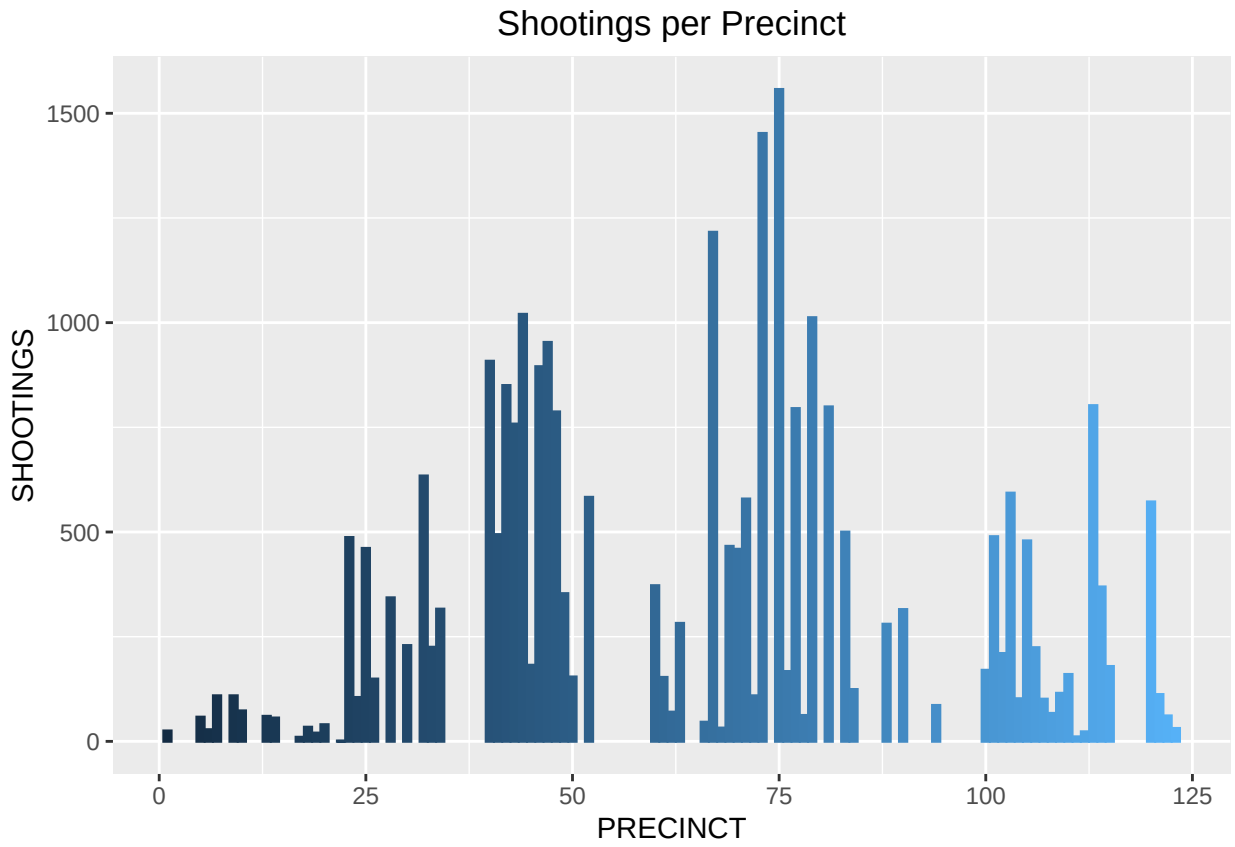
- https://www.nytimes.com/2022/12/14/nyregion/nypd-pay-work-costs.html

It would be worth analyzing how the distribution of the available patrol officers compares to the historical rates of shootings in each borough. One could argue that aligning the percent of officers in each borough with the percent of shootings in each borough might result in fewer shootings overall.

Here's how many shootings have been recorded in each precinct over the duration of the available data:

```
Shootings_by_Precinct <- Shootings %>%
  select(PRECINCT)

Shootings_by_Precinct <- Shootings_by_Precinct %>%
  count(PRECINCT) %>%
  rename(SHOOTINGS = n)
```

```
Shootings_by_Precinct %>%
  ggplot(aes(x = PRECINCT, y = SHOOTINGS, fill=PRECINCT)) +
  ggtitle("Shootings per Precinct") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_col(aes(color = PRECINCT), show.legend = FALSE)
```

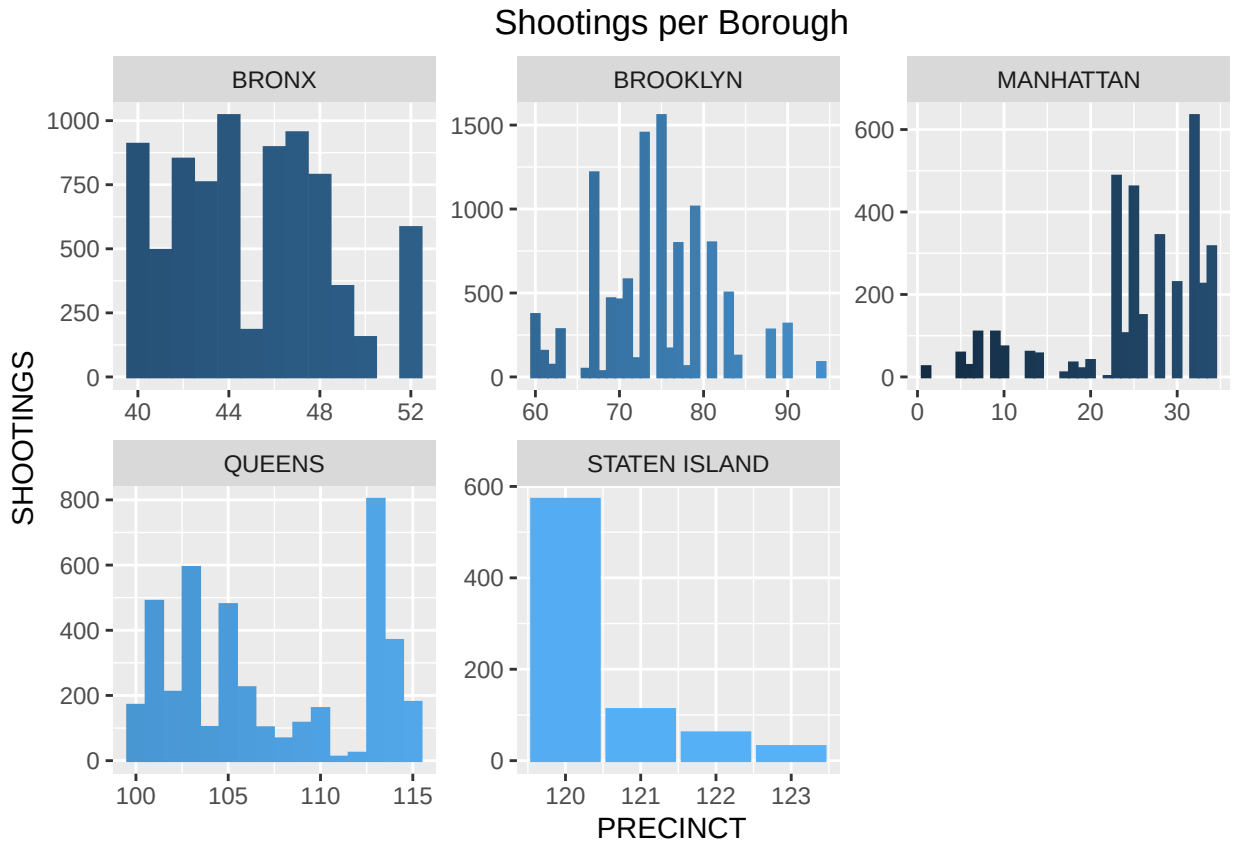


Here's another view of the same data, with each borough's distribution scaled to the same degree:

```
Shootings_by_Boro <- Shootings %>%
  select(BORO, PRECINCT)
```

```r
Shootings_by_Boro <- Shootings_by_Boro %>%
  group_by(BORO, PRECINCT) %>%
  count(BORO, PRECINCT) %>%
  arrange(PRECINCT) %>%
  rename(SHOOTINGS = n)
```

```r
Shootings_by_Boro %>%
  ggplot(aes(x = PRECINCT, y = SHOOTINGS, fill=PRECINCT)) +
  ggtitle("Shootings per Borough") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_col(aes(color = PRECINCT), show.legend = FALSE) +
  facet_wrap(~BORO, scales = "free")
```



Brooklyn appears to have the most shootings, followed by the Bronx, Queens, Manhattan, and Staten Island. It also appears the majority of shootings occur within a subset of each boroughs' precincts (ex., in Staten Island, about 75% of all shootings occur in the jurisdiction of just 1 of the 4 precincts.)

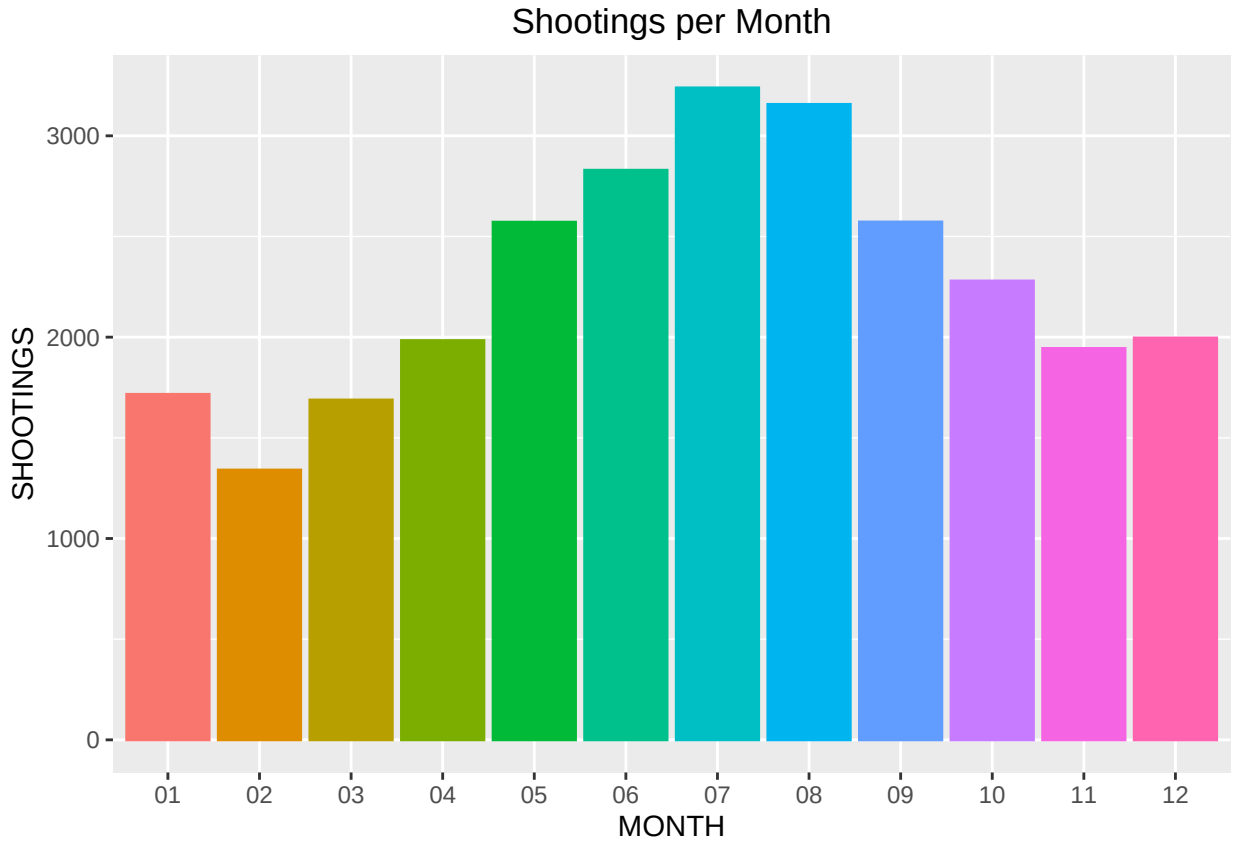
The populations of civilians within each borough/precinct are missing from the data set. Were the data made available, it would allow for the calculation of shootings per capita, which could further help inform any potential staffing reallocation. Of course, additional data showing the number of active duty officers in each boro/precinct at the time of each shooting would be useful in determining any staffing allocation changes as well.

Now that we know *where* the shootings tend to occur, let's see *when* they tend to occur.

Here are the number of shootings that have occurred each month:

```
Shooting_Months <- Shootings %>%
  select(OCCUR_DATE)

Shooting_Months <- Shooting_Months %>% mutate(MONTH = substr(Shooting_Months$OCCUR_DATE, 1, 2)) %>%
  count(MONTH) %>%
  rename(SHOOTINGS = n)

# Shooting_Months$MONTH <- as.integer(Shooting_Months$MONTH)
```

```
Shooting_Months %>%
  ggplot(aes(x = MONTH, y = SHOOTINGS, fill=MONTH)) +
  ggtitle("Shootings per Month") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_col(aes(color = MONTH), show.legend = FALSE)
```



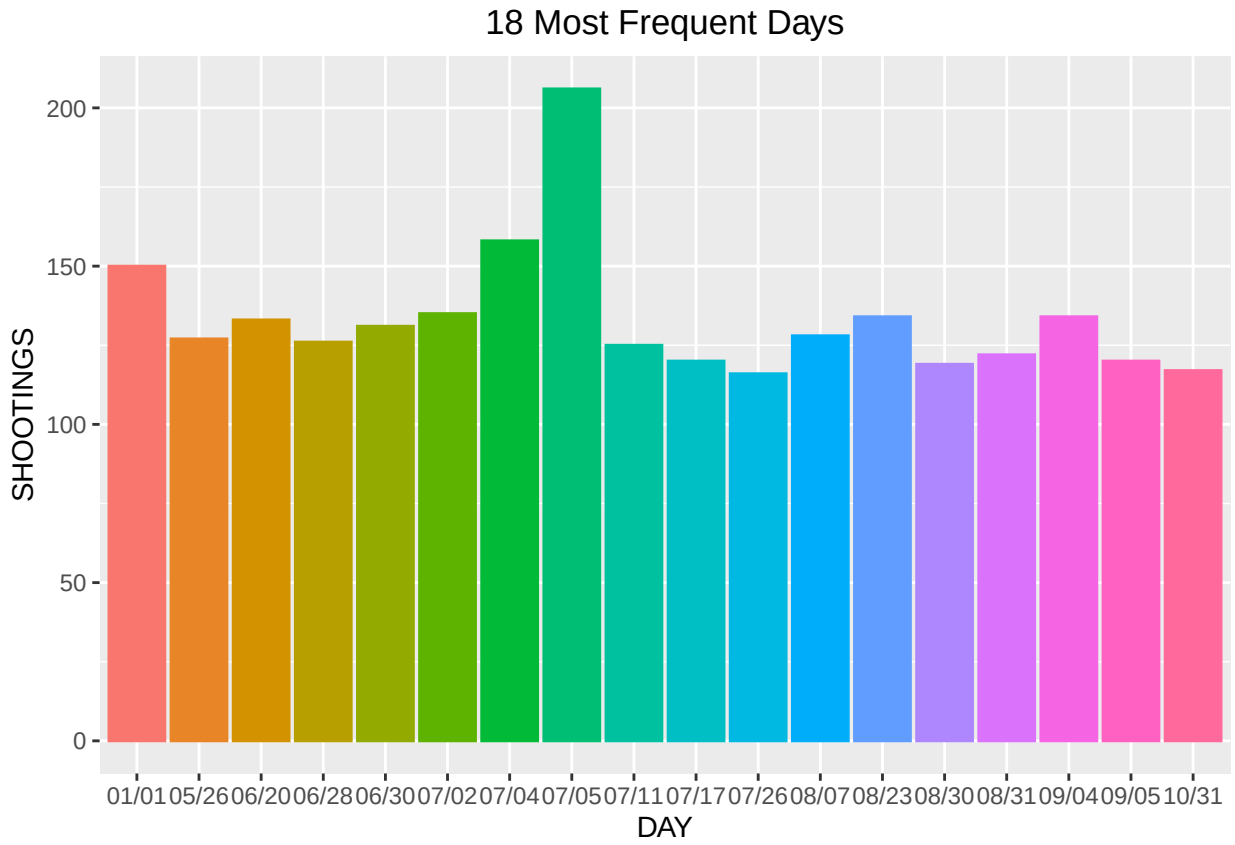
The trend of shootings over time is rather cyclical, with the least active month being February and the most active month being July. The months of December and January break this pattern as December has more recorded shootings than November and January more than February.

Here are the worst offending days:

```
Shooting_Days <- Shootings %>%
  select(OCCUR_DATE)
```

```r
Shooting_Days <- Shooting_Days %>%
  mutate(DAY = substr(Shooting_Days$OCCUR_DATE, 1, 5)) %>%
  count(DAY) %>%
  rename(SHOOTINGS = n)

Top_Shooting_Days <- Shooting_Days %>%
  filter(SHOOTINGS > 115)
```

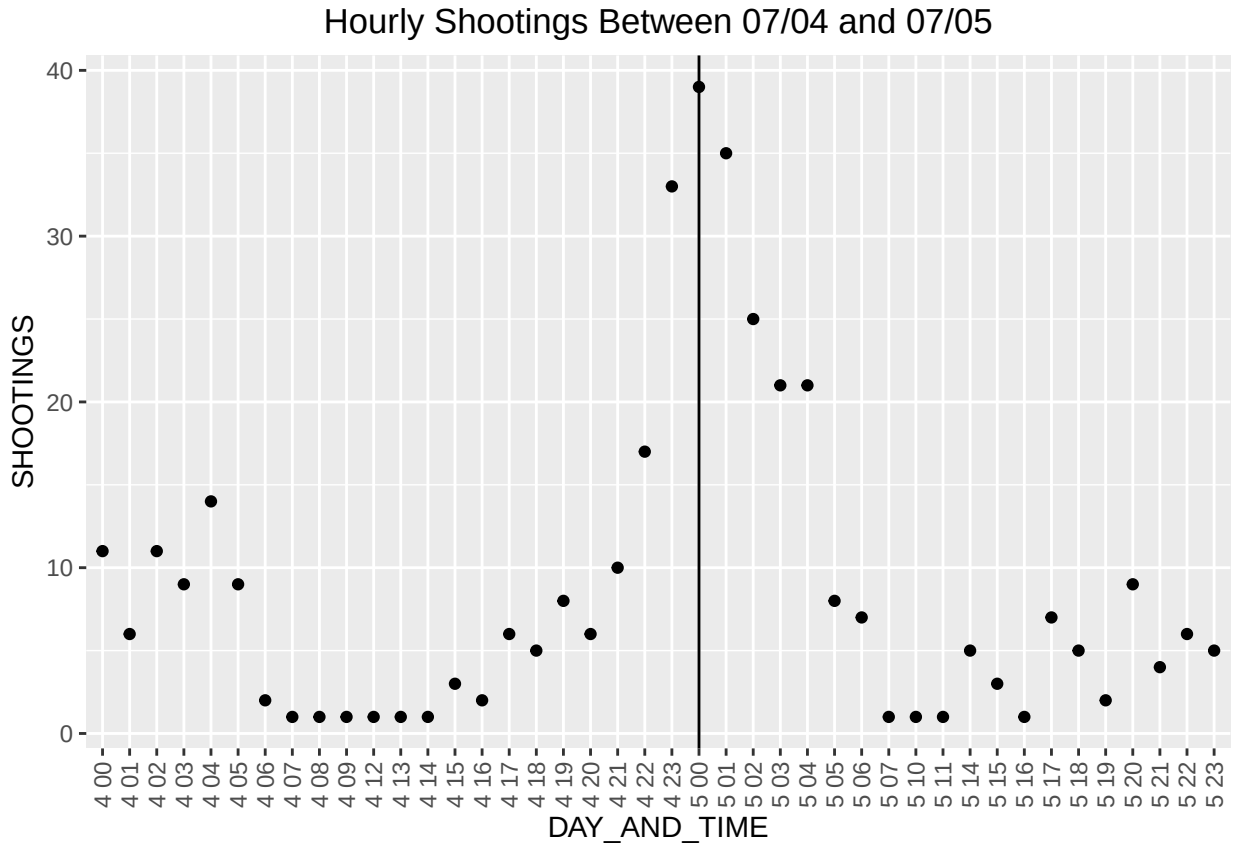
```r
Top_Shooting_Days %>%
  ggplot(aes(x = DAY, y = SHOOTINGS, fill=DAY)) +
  ggtitle("18 Most Frequent Days") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_col(aes(color = DAY), show.legend = FALSE)
```



Interestingly enough, the day with the most shootings is the day *after* the 4th of July (07/05), with the 4th of July itself being the day with the second most shootings. However, once we look at the times of each shooting, we see more accurately that the bulk of them happen between 12:00am and 4:00am on 07/05, which is effectively late into the night after the big holiday of the 4th of July:

```r
J45 <- Shootings %>%
  select(OCCUR_DATE, OCCUR_TIME) %>%
  filter(substr(Shootings$OCCUR_DATE, 1, 5) == "07/05" | substr(Shootings$OCCUR_DATE, 1, 5) == "07/04")

J45 <- J45 %>%
  mutate(DAY_AND_TIME = paste(substr(J45$OCCUR_DATE, 5, 5), substr(as.character(OCCUR_TIME), 1, 2))) %>%
  count(DAY_AND_TIME) %>%
```

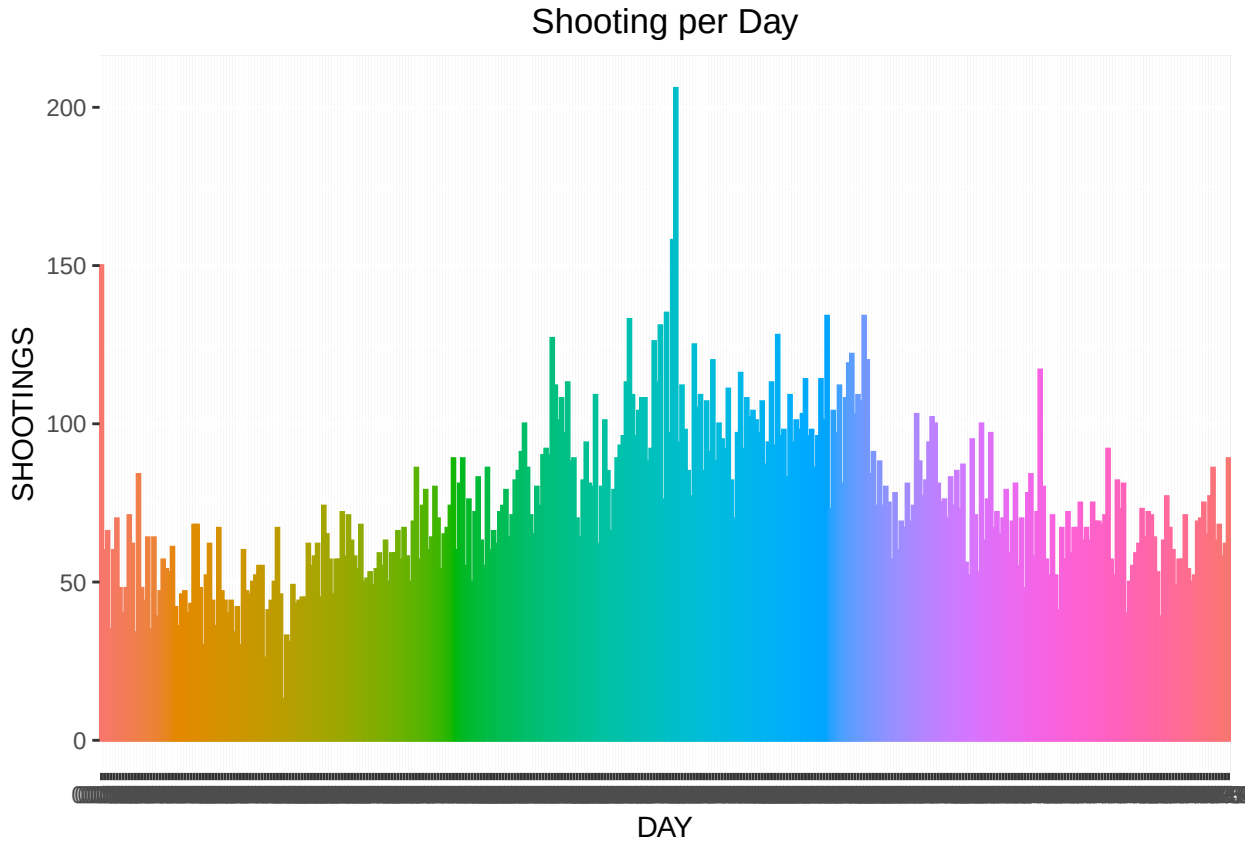
```
  rename(SHOOTINGS = n)

J45 %>%
  ggplot(aes(x = DAY_AND_TIME, y = SHOOTINGS, group = 1)) +
  ggtitle("Hourly Shootings Between 07/04 and 07/05") +

  geom_point() +
  geom_vline(xintercept = "5 00") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```



Other days with notably high levels of shootings include New Years Day and Halloween. These two holidays are of particular note since the monthly data suggests October and July have relatively few shootings as compared to the warmer months.

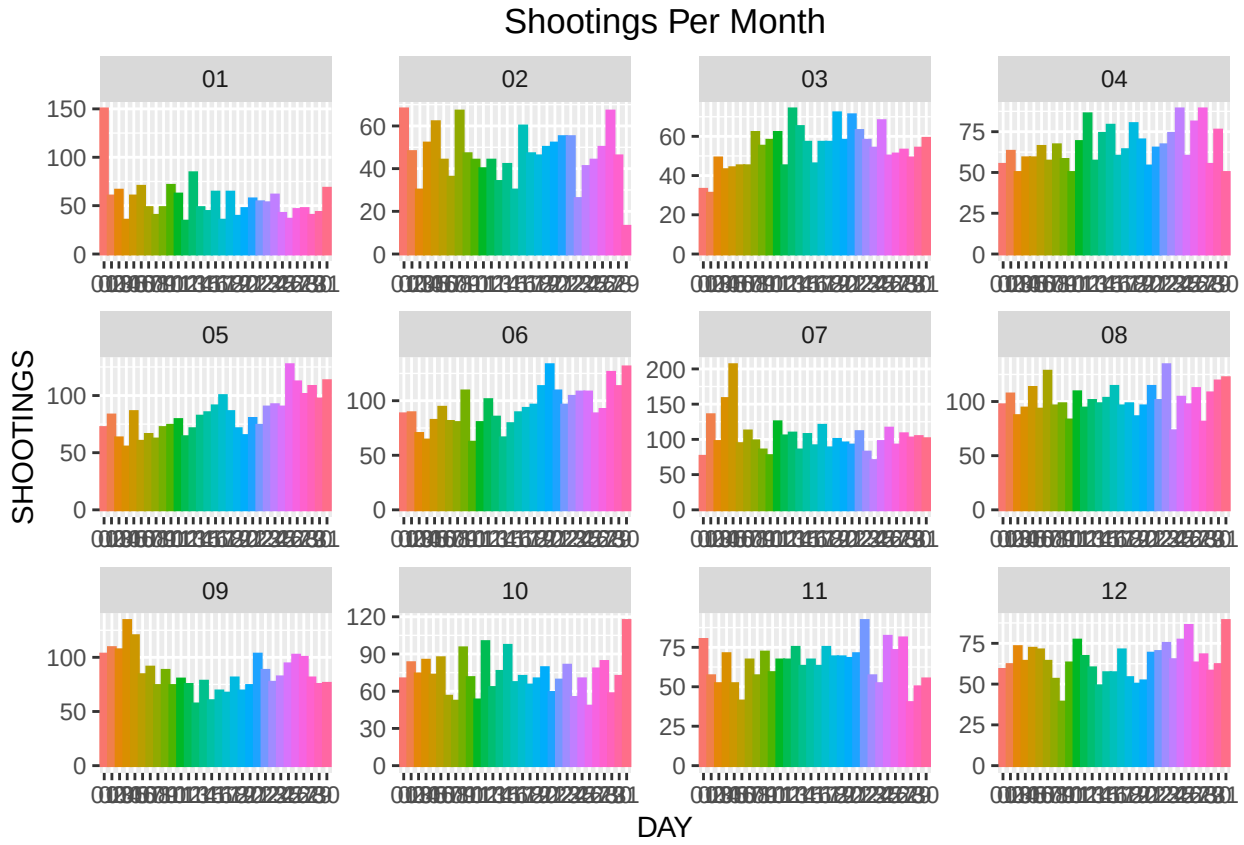
Now let's zoom out a little further and look at the shootings per each of the 366 days on the calendar:

```
Shooting_Days %>%
  ggplot(aes(x = DAY, y = SHOOTINGS, fill=DAY)) +
  ggtitle("Shooting per Day") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_col(aes(color = DAY), show.legend = FALSE)
```




```
Shootings_Per_Month <- Shootings %>%
  select(OCCUR_DATE)

Shootings_Per_Month <- Shootings_Per_Month %>%
  mutate(DAY = substr(Shootings_Per_Month$OCCUR_DATE, 4, 5),
         MONTH = substr(Shootings_Per_Month$OCCUR_DATE, 1, 2)) %>%
  group_by(MONTH, DAY) %>%
  count(MONTH, DAY) %>%
  rename(SHOOTINGS = n)
```

```
Shootings_Per_Month %>%
  ggplot(aes(x = DAY, y = SHOOTINGS, fill=DAY)) +
  ggtitle("Shootings Per Month") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_col(aes(color = DAY), show.legend = FALSE) +
  facet_wrap(~MONTH, scales = "free")
```

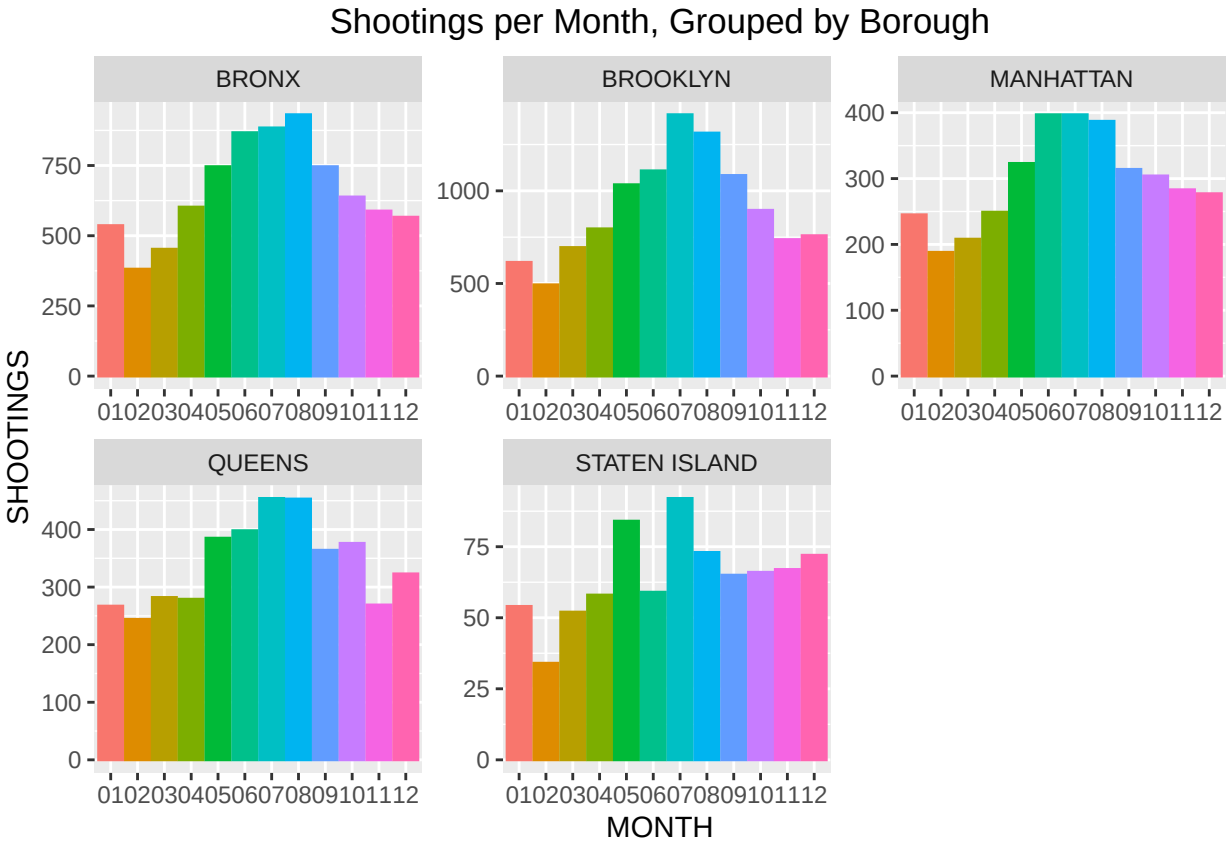


Within each month, there doesn't seem to be much of a pattern as to the rates of shootings. The only exceptions might be May, June, and September. A potential explanation for the general increase in the former 2 months and the general decrease in the latter could be not only attributed to weather, but also the school year. One more specific anomaly to observe is February 29th. Since this date only occurs once every 4 years, it's not too surprising that it's far less represented in the data.

Now let's look at the number of shootings that occur in each borough on a monthly basis:

```
Shooting_Months_Boros <- Shootings %>%
  select(OCCUR_DATE, BORO)

Shooting_Months_Boros <- Shooting_Months_Boros %>% mutate(MONTH = substr(Shooting_Months_Boros$OCCUR_DA
  group_by(BORO, MONTH) %>%
  # count(BORO, MONTH) %>%
  count(BORO, MONTH) %>%
  rename(SHOOTINGS = n)
```

```
Shooting_Months_Boros %>%
ggplot(aes(x = MONTH, y = SHOOTINGS, fill=MONTH)) +
ggtitle("Shootings per Month, Grouped by Borough") +
theme(plot.title = element_text(hjust = 0.5)) +
geom_col(aes(color = MONTH), show.legend = FALSE) +
facet_wrap(~BORO, scales = "free")
```

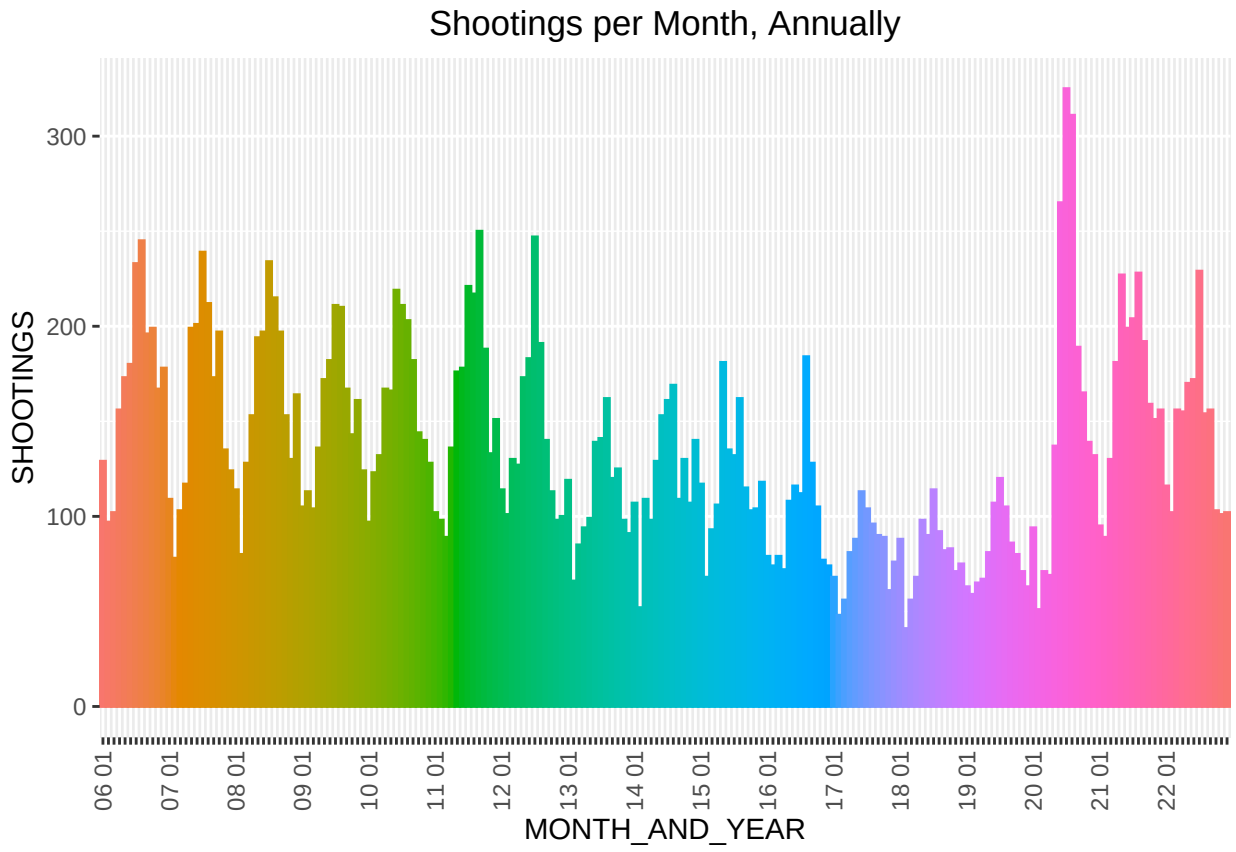


For some reason, there appears to be a sharp drop off in shootings in the month of June in Staten Island, while the other 4 borough see an increase in the same month. This would lead one to wonder if the data is fully accurate, or if there's a valid reason for this deviation.

One final thing we'll analyze here in addition to the aggregate monthly data is the annual monthly data:

```
Shootings_Yearly <- Shootings %>%
  select(OCCUR_DATE)

Shootings_Yearly <- Shootings_Yearly %>%
 mutate(MONTH_AND_YEAR = paste(substr(Shootings_Yearly$OCCUR_DATE, 9, 10), substr(Shootings_Yearly$OCCU
  arrange(MONTH_AND_YEAR) %>%
  count(MONTH_AND_YEAR) %>%
  rename(SHOOTINGS = n)
```

```
Shootings_Yearly %>%
ggplot(aes(x = MONTH_AND_YEAR, y = SHOOTINGS, fill=MONTH_AND_YEAR)) +
ggtitle("Shootings per Month, Annually") +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
      plot.title = element_text(hjust = 0.5)) +
scale_x_discrete(labels = function(x) ifelse(substr(x, 4, 5) == "01", x, "")) +
geom_col(aes(color = MONTH_AND_YEAR), show.legend = FALSE)
```

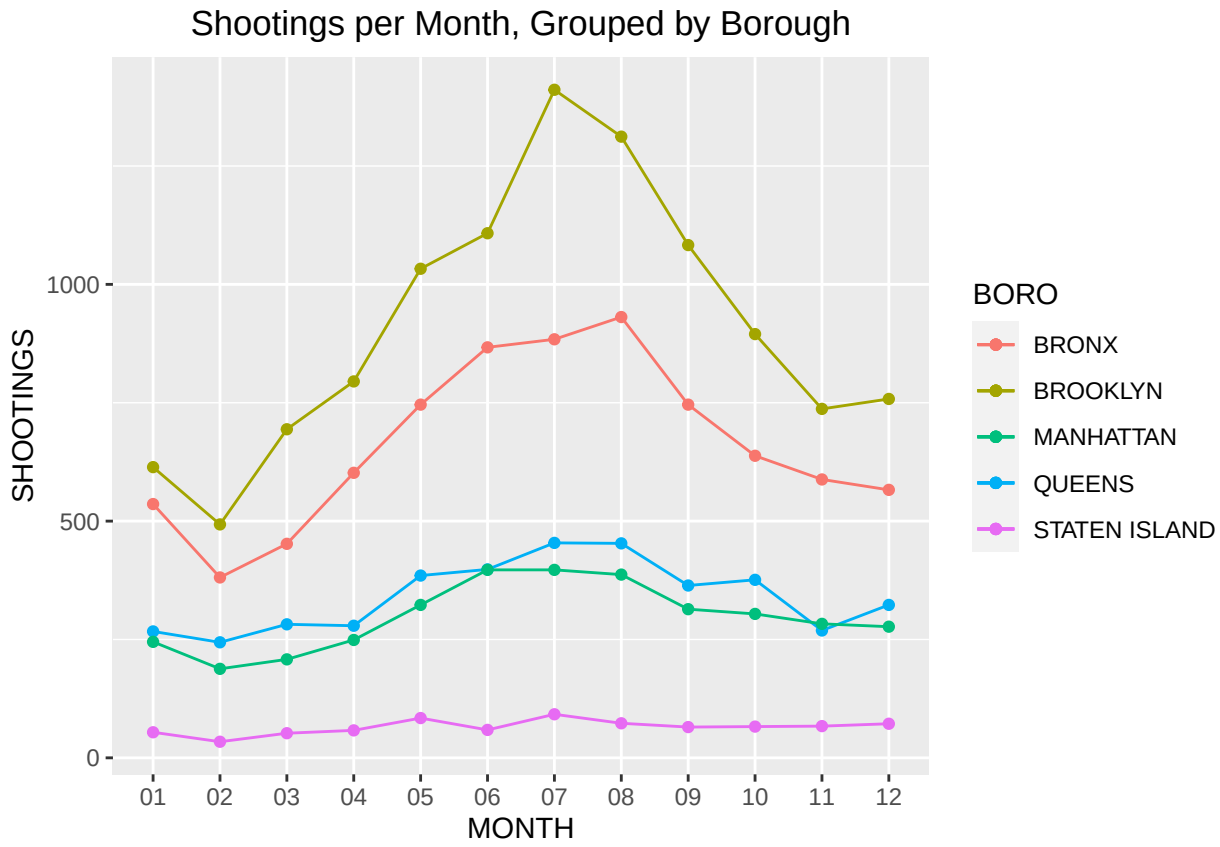


Over time, the number of shootings was trending down until there was a dramatic resurgence in the summer of 2020.

Putting it all together, here are the trend lines of the same data from above:

```
Monthly_Shootings_Brooklyn <- Shooting_Months_Boros %>% filter(BORO == "BROOKLYN")
Monthly_Shootings_Queens <- Shooting_Months_Boros %>% filter(BORO == "QUEENS")
Monthly_Shootings_Manhattan <- Shooting_Months_Boros %>% filter(BORO == "MANHATTAN")
Monthly_Shootings_Bronx <- Shooting_Months_Boros %>% filter(BORO == "BRONX")
Monthly_Shootings_Staten_Island <- Shooting_Months_Boros %>% filter(BORO == "STATEN ISLAND")

Shooting_Months_Boros %>%
  ggplot(aes(x = MONTH, y=SHOOTINGS, group = 1, color = BORO)) +
  ggtitle("Shootings per Month, Grouped by Borough") +
  theme(plot.title = element_text(hjust = 0.5)) +

  geom_line(data = Monthly_Shootings_Brooklyn) +
  geom_point(data = Monthly_Shootings_Brooklyn) +

  geom_line(data = Monthly_Shootings_Queens) +
  geom_point(data = Monthly_Shootings_Queens) +

  geom_line(data = Monthly_Shootings_Manhattan) +
  geom_point(data = Monthly_Shootings_Manhattan) +

  geom_line(data = Monthly_Shootings_Bronx) +
  geom_point(data = Monthly_Shootings_Bronx) +
```

```
  geom_line(data = Monthly_Shootings_Staten_Island) +
  geom_point(data = Monthly_Shootings_Staten_Island)
```



Some interesting observations in this data are:

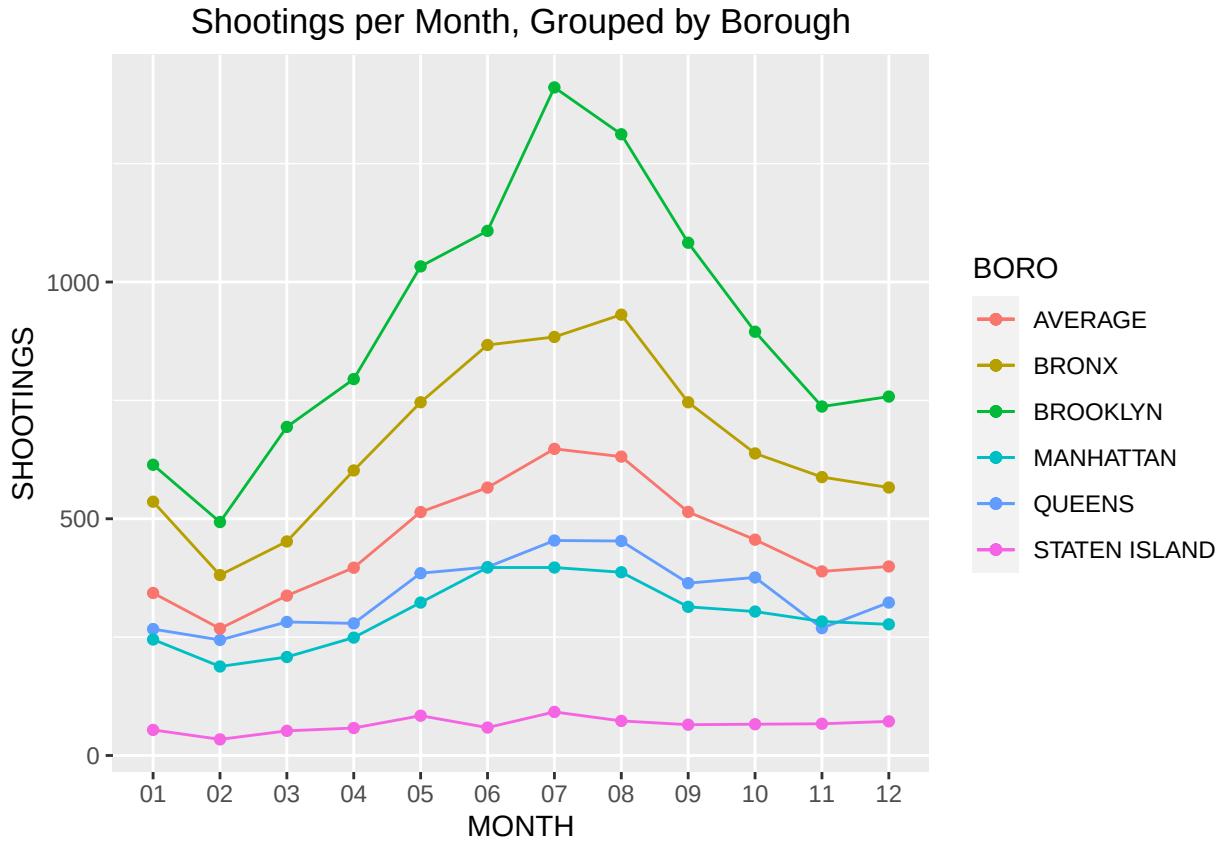
- Queens and Manhattan have nearly identical shooting numbers for the month of June.
- For 11 of the months, Queens has had more shootings than Manhattan. But for some reason, Manhattan has seen more shootings in the month of November than Queens.
- Brooklyn has a dramatic uptick in shootings in the month of July.

To model the data, we'll look at the average shootings in each month:

```
#   Average_per_Month <- Shootings %>%
#   select()
#   group_by(MONTH) %>%
#   mutate(avg = mean(SHOOTINGS))
# # Shooting_Months_Boros <- Shooting_Months_Boros %>% mutate(pred = predict(Average_per_Month))

Average_Per_Month <- Shooting_Months %>%
  mutate(SHOOTINGS = SHOOTINGS/5) %>%
  mutate(BORO = "AVERAGE")
```

```
Shooting_Months_Boros %>%
  ggplot(aes(x = MONTH, y=SHOOTINGS, group = 1, color = BORO)) +
  ggtitle("Shootings per Month, Grouped by Borough") +
```

```
  theme(plot.title = element_text(hjust = 0.5)) +

  geom_line(data = Monthly_Shootings_Brooklyn) +
  geom_point(data = Monthly_Shootings_Brooklyn) +

  geom_line(data = Monthly_Shootings_Queens) +
  geom_point(data = Monthly_Shootings_Queens) +

  geom_line(data = Monthly_Shootings_Manhattan) +
  geom_point(data = Monthly_Shootings_Manhattan) +

  geom_line(data = Monthly_Shootings_Bronx) +
  geom_point(data = Monthly_Shootings_Bronx) +

  geom_line(data = Monthly_Shootings_Staten_Island) +
  geom_point(data = Monthly_Shootings_Staten_Island) +

  geom_line(data = Average_Per_Month) +
  geom_point(data = Average_Per_Month)
```



Not too surprisingly, the average trend line tracks the shape of the "Shootings per Month" bar graph above. But now overlaying this trend line with the individualized boroughs' monthly data lets us see just how disparately the different boroughs are affected.

In summary, the rates of shootings seems to be cyclical. There is generally an increase from February to July before a general decrease from July to November, with another increase/decrease cycle from November to February. Brooklyn experiences the most shootings by far, peaking in the month of July. If one was to

believe increased police patrolling would lead to a decrease in shootings, the data would suggest the rate of patrolling should change accordingly.

Some potential sources of bias in the data include:

- These are only reported shootings. There's almost certainly a non-zero amount of shootings thaat have gone unreported.
- There could be missing data points in the provided data. To mitigate this, incidents where any of the OCCUR_DATE, OCCUR_TIME, PRECINCT, or BORO data points were missing were removed from the analysis.
- My own bias suggests more police could lead to reduced shootings. However, further analysis would be warranted to prove this relationship is truly causal. That said, the conducted data analysis itself here does not necessarily have this bias associated with it. This particular bias would be more related to actions taken as a result of the data presented.